

Resilience in identifier design, and some thoughts on granularity  
David Shetland, Thomas Bruce  
Legal Information Institute, Cornell Law School

The Legal Information Institute (LII) has a great deal of experience with processing the United States Code (USC), as well as US Supreme Court opinions and various other legislative and case law resources. Here we address two closely related questions that confront the XML and Legislation Workshop associated with Jurix 2007, having to do with our use of identifiers and the means by which we determine granularity.

The USC consists of about 50,000 sections containing the statutes as codified, together with about 8,000 section-size container elements like Chapter. All authoring and revision is done by the US Congress and its Office of the Law Revision Counsel before the LII receives it. It is transmitted in a specialized, highly modal encoding used for typesetting.

Decisions about appropriate granularity and identifier schemes -- which in any case are closely related -- have been driven by the realities and contingencies of an automated filtering process used to create an XML version of the US Code from this typesetting data. This is perhaps an easier task than supporting the full life-cycle of drafting, passage, and promulgation, but suffers from noisier and occasionally mis-structured data, over which we have little initial control. Some aspects of that processing system, of particular interest to identifier design and usage, are listed here:

(1) The initial XML schema emphasizes retention of source data, much of which will not be understood well at the outset, or perhaps ever, but all of which is considered a potentially valuable resource.

(2) The system, including the XML, supports an incremental improvement approach to citation reference detection and associated link generation. Specifically, the system needs to determine whether a potential cross-reference proposed by the automatic linker is valid.

(3) Xpath specifications can be safely used for structured access to elements without identifiers within small structural units, such as certain paragraph fragments and references. This would not be true of larger units.

(4) XML fragment size is currently that of USC section, which allows for efficient file sizes (one fragment per file). The file sizes range over three orders of magnitude, corresponding to variations in USC section bulk and complexity.

(5) Element instance identifiers are used. These distill and normalize primary identifying data, and represent deliberate compromise between the extremes of mindless abstraction and mindless recitation. Thus, instead of the literal text citation section number of 12\_USC\_1749bbb-10a, we see it padded out to usc\_sec\_12\_00001749-bbb010a for ease of collation, but not converted into a hash code requiring data base reflection to yield any semantic value.

Data-set independent identifiers should be added to the model to allow for tracking of arbitrary movement of content. We have yet to add point-in-time historical tracking, and we have no role in the editing process. A system-defined identifier facility will be needed at some time, but has not yet been implemented. When it is used, it will answer the single primary ID expectation. But other identifiers, including data-derived ones such as we now employ as primary, would still be needed for portability, archiving, efficiency of associating meta-data with structural units, convenience of programming, and compatibility with real-world, legacy identifier systems. Existing print-based citation systems and structural labels that have become terms of art within the profession (eg. 501(c)(3) nonprofit ) are particularly important in this respect.

These factors have led us to designs in which semantically-neutral identifiers are used internally, and semantically-laden identifiers provided externally by means of a software translation layer. This allows association of a particular fragment with multiple identifier schemes. While this is a less important capability for legislation than it is for caselaw, where a particular case may have many citations over its lifetime, it is nevertheless extremely useful.

#### Granularity

For us, granularity has often been a matter of specifying the smallest structural unit that we need to make externally addressable. We have deliberately blurred the dual nature of identifiers as labels. At times they are pinpoint destinations within a text stream; at others, they are handles by which structural units can be retrieved. The smallest unlabeled block of text that we treat as a container is the paragraph. Were we to support the drafting and revision of legislation, we might need finer granularity. But it seems to us that there is a reasonable lower limit; rather than assigning identifiers to bytes or tokens, it makes sense to turn fine-grained editing over to subsystems that take (eg.) paragraphs as input and produce altered paragraphs as output.

#### Summary

The task of presenting statutory text derived from a legacy data format, in a manner that retains authority, is far different from the task of defining a new approach to authoring the source data. Techniques and devices for discovery of existing material in a data set ( reflection , etc.) as well as for resolution of addressing ambiguities are needed by both kinds of support systems. But a variety of identifier schemes may be expected in support of the different systems. In fact they will co-exist within any system supporting migration from legacy data. Any single-identifier mandate must be considered external to the mandates inherent in the base data sets, which must be allowed their own identifier schemes, if only recorded as part of the routine metadata.